Controlling Bad Behavior in Online Communities: An Examination of Moderation Work

Research-in-Progress Paper

Aiden McGillicuddy
PwC New Zealand
113-119 The Terrace, Wellington 6011
New Zealand
aiden.mcgillicuddy@nz.pwc.com

Jean-Grégoire Bernard
Victoria University of Wellington
23 Lambton Quay, Wellington 6140
New Zealand
jean-gregoire.bernard@vuw.ac.nz

Jocelyn Cranefield
Victoria University of Wellington
23 Lambton Quay, Wellington 6140
New Zealand
jocelyn.cranefield@vuw.ac.nz

Abstract

A wide range of behavior may be seen as destructive to online communities. Yet behavior that is 'bad' in one community may be celebrated in another. The work of community maintenance is therefore strongly contextual, involving complex choices due to differing norms, community cross-membership, and the need to invoke fairness. The experienced, "lived in" work of moderators; how they enact norms and make choices about social maintenance, remains poorly understood. Our study addresses this gap, using a negotiated order lens. We employed netnographic techniques, analyzing online interviews with moderators of sub-communities in Reddit, and records of critical incidents. We find that moderators are intuitive prosecutors who draw on a variety of logics to accomplish their work. Controlling bad behavior, articulating and enforcing norms is, therefore, a collective accomplishment through which moderators make choices, create a jurisprudence record, and reconcile nested community norms in the maintenance of social order.

Keywords: Online communities, Work practices, Social dynamics, Social norms, Theory building

Introduction

Bad behavior abounds on the Internet: spamming wastes time, trolling can lure people into pointless arguments, and flame wars can polarize communities. Yet rule-breaking behavior in one community may be entirely acceptable in another. While trolling on Wikipedia leads to the spread of misinformation and the disengagement of editors, in the case of the community 4chan, trolling is an accepted part of culture (Bullard & Carter, 2013). The task of policing member behavior typically falls to community moderators. In order to ensure the effective functioning and survival of communities, moderators must sanction inappropriate behavior and police norms in a way that is seen as fair. However, due to the diversity of online community research and the widely varying nature of online communities, the role of moderators in maintaining community norms is not well understood. In particular, we lack a high-level understanding of the experiences involved in, and decision-making principles underlying their work.

Our exploratory study sought to answer the question, *how do individuals in roles responsible for online community maintenance make choices about how to moderate and sanction the behavior of community members?* We used a negotiated order lens (Barley & Kunda, 2001; Bechky, 2011) to identify how community norms and principles were interpreted and applied in practice by moderators of different communities in Reddit.

The purpose of this paper is to report on the preliminary findings of our study in progress. In this paper our emphasis is on opening up the black box of moderation work by rendering our informants' experience transparent and previewing the constitutive elements of our theorizing process, rather than proposing a complete, polished theory about the dynamics of moderation work. Our main contribution with this paper is to reveal the variegated contexts in which moderation work takes place, and to highlight the tensions and dilemmas involved in moderation work. The nature of our contribution lies in reporting rich data that can't be readily explained by our field's current arsenal of theories about online communities, and thus we aim to stimulate future theory development about the nature of moderation work (Avison & Malaurent, 2014; Henfridsson, 2014; Markus, 2014). In the sections that follow we firstly review the literature about bad behavior online, its relationship to community norms, and the roles of moderators. We then explain our theoretical lens before describing our research method and outlining our preliminary findings. We end the paper by conjecturing about the possible meanings of our findings for a number of theories about online community maintenance in the digital age.

Theoretical Background

Bad Behavior in Online Communities: What is it and Why Does it Matter?

Since Dibbell (1993) examined cyber rape in the LambdaMOO community, many scholars have focused on understanding, managing, and preventing the phenomena of 'bad' online behavior. Flaming, trolling, griefing, cyber bullying, doxing and spamming are typically seen as destructive to a community and its users (e.g., Bergstrom, 2011; Binns, 2012; Campbell et al., 2009; Chesney et al., 2009; Kiesler, Kraut, Resnick, & Kittur, 2012; Lampe, Zube, Lee, Park, & Johnston, 2014). Online communities operate within sets of norms; implicit or explicit rules and guidelines (Kollock, 1994), linked with established and expected behavior patterns (Warren, 2003), but what is unacceptable behavior in one online community may be the norm in another (Binns, 2012; Dibbell, 1993; Kiesler et al., 2012). As a result, entering a new community may lead to norm violation. Anonymous users are more likely to violate norms because they do not face social sanctions (Kiesler et al., 2012). Further, if a user does not identify with a community they will more readily behave in a way that reflects individual attitudes (Goldspink, 2010). However, even community members who understand the norms may not comply with them: they may contest norms or display behavior reflecting their expectations of what they think should be the norms (Kiesler et al., 2012).

While 'bad' behavior varies according to context, a common concern is the externalities invoked by reactions to problematic behavior. If users over-contribute it may overload users' capacity to take in the information. Lampe and Johnston (2005) found that after a critical mass was reached, contributions were discouraged by the rate of contributions from others. Also, if there are many unworthy contributions, it is difficult to gain useful information (Kollock, 1994). However, in some circumstances, bad behavior may be beneficial. In the case of constructive deviance (Vadera et al., 2013; Warren, 2003) behavior that departs from local norms may nonetheless benefit the community by aligning it with societal norms.

The Moderator's Role in Maintaining Community Behavior

Online community users learn how to behave through experience, observation, and feedback (Lampe & Johnston, 2005). The judgments of others are often the best indicator of what messages are worth paying attention to (Lampe & Resnick, 2004). The job of making decisions surrounding the appropriateness of online behavior and how to respond to it typically falls to volunteer community moderators. Moderators ("mods") perform diverse duties to remove and/or reduce the impact of negative messages, such as screening, labeling, moving or removing inappropriate messages, and downgrading content through techniques such as "disemvoweling"; removing vowels to diminish readability (Kiesler et al., 2012).

Methods that diminish the impact of a speaker rather than silencing them (such as redirecting comments to an off-topic forum) are likely to be met with less resistance (Kiesler et al, 2012). Nonetheless, mods are may throttle or gag users who show signs of creating significant damage.

Moderation work is typically performed within a larger system of complementary systems such as reputation systems, reversion tools, quotas and technical limitations on the creation of new accounts (Kiesler at al., 2012) but is nonetheless time-consuming and at times emotionally exhausting (Butler, Sproull, Kiesler & Kraut, 2002). It is also complex work: for the community's continuation, it important to invoke a sense of procedural justice, the belief that sanctions are given fairly and consistently (Kiesler et al., 2012) and to balance forms of social control with encouragement activities (Butler et al.; 2002). In large online communities, moderation is a substantial job that may be distributed and layered. For example, Slashdot allows thousands of users to moderate comments[1] while employing meta-moderation; moderating of the fairness and accuracy of this moderation (Lampe & Resnik, 2004; Poor, 2005). It appears that moderators are motivated to do this voluntary social maintenance work for a combination of informational, social, visibility and altruistic benefits (Butler et al., 2002; Bateman, Gray, & Butler, 2011).

It is common to see moderation explained as a sophisticated set of regulation activities, or human system (e.g. Lampe & Resnik, 2004; Kiesler et al., 2012; Poor, 2005). Such accounts of moderation are abstract and procedural rather than experiential in nature. While such studies suggest the complexity of moderation, their purpose is to outline community systems and/or design principles rather than to convey the human nature of moderating work. At present we know more about the 'what' of moderation than how the role is performed and lived in. Given that communities face a considerable challenge in getting people to devote the required time, care and attention for maintenance activities (Butler et al. 2002), it is critical to better understand the nature of moderation work.

Negotiated Order as a Lens for Examining Online Moderation

Our study of moderation employs the theoretical lens of negotiated order. Negotiated order theory (originating with Maines; 1978, 1982; Strauss, 1963; Strauss, Schatzman, Bucher, Erhrlich and Sabshin, 1963; 1978) rejects the view that organizations are inherently stable entities. Instead, it views order as a complex and ongoing social accomplishment that involves continous negotation. In contrast to classical organizational theory, which is typically based around abstract structures, research in the negotiated order tradition is concerned with how people live, define, negotiate, and perform, their work in practice (Bechky, 2011; Hallett, Shulman, Fine and Adler, 2009). This body of theory takes the view that people's response to situations is interwoven with their interpretations of these situations (Bechky, 2011) and that power is not a stably bestowed entity, but is dynamically distributed between stakeholders (Magnusson and Nilsson, 2013). For example, Strauss, Schatzman, Bucher, Erhrlich and Sabshin's (1963) study in a hospital reveals how the hospital rules and structures are "interwoven into the working arrangements of doctors, nurses, and administrators" who alternately draw on and ignore them as they go about their work (Bechky, p. 1160). The recursiveness involved in negotiating social order is strongly emphasized by this theoretical lens: Dokko (2012) describes negotiated order theory as providing "a model of social order that is maintained through a recursive relationship between a structural context, a more proximate negotiation context, social interactions, and interaction outcomes" (Dokko, 2012; p.686). The negotiated order lens also calls for attention to the values and logics of action enacted by role occupants in organizations; for instance, Becker (1973) examined the performances of moral entrepreneurs in creating and enforcing social norms.

The negotiated order lens has previously been used in the IS field; for example, to examine how technology affects work and organizations (e.g., Barley, 2015; Lamb & Kling, 2003; Kling, 1978; Magnusson and Nilsson, 2013). Dokko (2012) has applied negotiated order theory to explain the trajectory of technological evolution through the role of technological communities. We elected to apply the theoretical lens of negotiated order in our analysis of moderation work owing to its focus on context and social interaction. This makes it suitable for examining Reddit moderation work, which is social situated, highly distributed and performed in a voluntary capacity across a considerable variety of online community settings, and how this work is impacted on by the social and contextual demands of online communities, and the technology available.

[1] https://slashdot.org/moderation.shtml

Methods

Research Design and Case Selection

Our inductive, theory development approach relies upon a comparative design of maintenance work in online communities hosted on Reddit, a global community platform (Barley & Kunda, 2001; Bechky, 2011; Gioia, Corley, & Hamilton, 2013). As of May 2016 Reddit hosts over 11,000 different communities, called subreddits, each having its own topic and purpose. Moderators, or "mods," perform the work of maintaining the community, and of particular interest for this study, of defining norms and sanctioning bad behavior. We follow the advice of Barley & Kunda (2001), by adopting a "within-family" design that allows "comparing and contrasting two or more lines of work that bear a strong family resemblance" (p.85). The value in choosing Reddit for such a design is that all communities adhere to a set of global norms (e.g., "reddiquette," content policy), offer a basic set of affordances to users (e.g., submit posts, upvote content, comment, reputation system), and provide a toolkit to aid moderators in their maintenance work. Yet, the communities also differ in the content they host, the demographics and size of their audience, and the local norms they follow. The interaction of these similarities and differences should provide the background necessary to surface the variants in how moderators control bad behavior.

In accordance with the comparative design chosen, we selected a sample of 21 communities (distinct subreddits) along several theoretical dimensions: We selected communities of varying visibility, as measured by their status as "default subreddit," which are featured on the home page of Reddit.com. We also selected for variance in size, as measured by subscriber count, and topic and purpose, by seeking communities that curate political, scientific, lifestyle, and entertainment content. Further, we sought variation in the degree and formality of moderation activity displayed by the community, taking into account frequency and amount of moderators' intervention and the extent to which norms were codified.

In each of the 21 communities identified, a moderator was identified. We purposefully sampled active moderators on a theoretical basis, by aiming for a wide range of background: involvement in more than one community, varying seniority, ranking, and Karma score, a reputation signal. Out of 21 mods invited to participate, 12 began the interview, 2 provided partial answers, and 6 answered all questions. An additional round of at least 20 interviews with moderators is planned to extend and further develop the preliminary findings we present in this paper.

Data Collection

This research has employed two data qualitative collection techniques drawn from netnography (Kozinets, 2015): (a) online interviews with moderators as a primary source of data and (b) online observation and records of critical incidents in the communities they moderated as a secondary source of data.

In order to gain access to the moderators in the Reddit environment (where anonymity is required), certain steps of cultural immersion were undertaken (Kozinets, 2015). To establish proof of his identity as a researcher, the first author set up a university-hosted web page describing the research and containing a photo of him holding a small sign, handwritten, with his Reddit username. He then initially approached moderators over Reddit's private messaging platform. The first author has been a long time active member of several Reddit communities, and thus was able to adapt his approach to the culture of each community. After rapport was established, participants were given the choice of conducting the interview over a medium of their preference, such as email or videoconference (Skype).

Privacy risk and lack of trust have been significant barriers to gaining access to participants so far. Mods who declined taking part in confidential interviews typically cited privacy concerns as their reason for not participating. Moderators are often part of a wider team of moderators, and the answers they give might result in repercussions if they are identified. Moderators can also be unwilling to divulge into sensitive topics as it reflected on them or a co-moderator negatively. Because the interviews were being conducted over a platform where anonymity is guaranteed by default, providing evidence of our role and intentions was difficult. Even when proof of this was provided, some participants doubted our intentions.

While online interviews do not provide the richness and spontaneity of face-to-face interviews, they still provide a convenient way to access hard-to-reach participants as well as to collect emic data about rationale and meaning of choices (Kazmer & Xie, 2008; Kozinets, 2015). An obvious limit of conducting interviews online is that some of the participants' deliberation is hidden from the interviewer, and there is a greater risk of break off from the interview. However, as a pragmatic solution to the anonymity assurances required by moderators of large, visible, and sensitive topic communities, it was assessed that the trade-off was acceptable. Future research efforts will aim at building further trust and rapport within Reddit communities, in order to get access to even richer accounts of moderation work.

The interview protocol was chunked in four sections. Questions focused on surfacing the understandings, experiences and perspectives of mods on four aspects of their work: (a) understanding of the community's norms, (b) the degree of formalization of the community's norms and reason for this, (c) how they enforce these norms, and (c) how user behavior that deviates from norms impacts the community. The participants were also asked to explain critical events in their subreddit's history which were then used as supplementary evidence. They were also encouraged to provide URLs relating to those incidents. All questions were written in an open-ended interview format in order to capture the rich descriptions. The protocol had been pilot-tested with moderators of a small geographic subreddit.

Data Analysis

To prepare for analysis, all interviews transcripts, archival record, and observations were imported into NVivo to create a case database for each community. We adopted Gioia et al.'s (2013) techniques of grounded data analysis: (1) inductively creating first order categories, (2) deductively creating second order themes from the 1st order categories, (3) and then aggregating dimensions from 2nd order themes.

The first order categories were identified without reference to the prior literature, any categories that stood out from the transcripts were created but no statements were coded. A large set of descriptive codes were developed by the first researcher to identify the expressed opinions, values and actions of moderators (this included moderation activities, reported moderation events, reports of user behavior, attitudes, rules, ways of working, use of technology, etc.). These codes were discussed amongst the researchers and a return to data was undertaken to ensure comprehensive coverage, an application of coding to statements, and saturation of descriptive themes. Finally, the transcripts were read through a third time to ensure all the participants answers were coded correctly. A typical outcome of this stage involve the parsing of short "chunks" from interview transcripts into low level codes that attempted as much as possible to keep our informants' terms and meaning intact. The statement "*I guess a direct contrast to us would be our "brother" sub, [subreddit X], which in contrast is quite lax with it's moderation. We specifically try and distance ourselves from being associated with them because our moderation styles are so different that users from their sub can be unwilling to adjust to how we do it here (or be really belligerent about it)*" was coded as *contrasting moderation styles from brother subreddits*. Some statements were double-coded, because they highlighted variations of the same theme. For instance, the statement "*[Subreddit X] run a tight ship with a focus on user comfort (in contrast to say [other subreddit] where the focus is content accuracy/quality)*" was coded as *user comfort logic* and *content quality logic*.

After reflecting on the earlier literature review and making conclusions about the inductively generated categories from the analysis, themes were deductively generated to make sense of the categories. For instance, the codes *user comfort logic* and *content quality logic* were grouped under the 2nd order theme of *orientation among control logics*. The code *contrasting moderation styles from brother subreddits* was grouped with other related codes under the theme *reconciling nested community norms*. The 2nd order themes that have been identified so far include: contribution cost, quality standards, community visibility, topic, orientation among control logics, sense of power, career stages, hiring moderators, labeling bad behavior, articulating norms, and reconciling nested community norms. The third step in our analytical approach was the creation of categories informed by the theoretical lens of negotiated order. At this point, 3 aggregate dimensions were created from the 2nd order themes and interrelated to create a model of how moderators deal with bad behavior: a community's demand for controlling bad behavior, a moderator's stance toward controlling bad behavior, and the collective accomplishment of controlling bad behavior. Two of these dimensions are based around the two key negotiating parties - the community, and the moderator, while the third combines codes that illustrate the collective accomplishment of negotiated order.

The themes and dimensions we report in this paper consist of the constitutive elements of our model of moderation work. Future data collection and analysis efforts will likely allow the identification of additional themes and dimensions that serve as the 'deep structure' of our nascent theory of moderation work (Gioia et al., 2013, pp.22). Further analysis about the dynamic interrelationships between themes and dimensions are still required to highlight the 'deep processes' of our nascent theory of moderation work. Because data collection is still ongoing for this study (an additional round of 20 interviews is underway), we have not yet engaged independent coders to code portions of the data and compare agreement, although this is a step that will be accomplished once the data collection effort is complete, in order to bolster our confidence in our interpretations.

Preliminary Findings

A range of communities has participated in the research project so far, ranging from 1.5-2.0 million to 1,000-20,000 subscribers. The density of norms for both large and small large communities ranged from high (extensive norms) to low (few) norms, and community foci ranged from serious discussion of serious topics to light-hearted fun. Table 1 outlines the characteristics of the communities of each moderator interviewed. Details are obscured to protect the privacy: the names of the community, subscriber count, age, topic, and ratio of moderators. The extent to which norms are codified was assessed based on the quantity and nature of codified norms that were listed on the subreddit page. A continuum exists from one extreme to the other where AskQuestions had the most codified norms and Townsville had the fewest.

Communities[2]	**Size (Subscribers)**	**Age**	**Topic**	**Codified Norms**	**Subscriber to mod ratio**
Townsville	20,000	< 5 years	Geographic	Almost none	1,500:1
Funny	500,000	> 5 years	Humor/satire	Almost none	12,000:1
DiscussIt	50,000	< 2 years	Discussion	A few norms	7,500:1
Gaming	20,000	> 2 years	Gaming	Medium level	500:1
Hobby	2,000,000	> 5 years	Special interest	Several	200,000:1
FunnyPictures	500,000	> 5 years	Humor/satire	Several	25,000:1
AboutWomen	500,000	< 5 years	Women's issues	Extensive	5,000:1
AskQuestions	500,000	< 5 years	Discussion	Extensive	11,000:1

Table 1. Community Profiles

We now present the constitutive elements of a nascent theory of the contingencies involved in accomplishing moderation work. We firstly consider why bad behavior is more or less common in online communities, by examining the demand for moderation work. Part of this demand is contingent on the cost required from members to make a contribution in terms of the quality standard through which contribution is assessed. Another part is contingent on the visibility of a community, and by the potential of a topic for controversy.

A moderator's stance toward their work will condition their response to bad behavior. Such stance is influenced by the moderator's ability and discretion to make compromises among logics that prescribe conflicting remedies to bad behavior. The moderator's sense of power will influence whether they see moderation as a fiduciary duty to the community or as a license to impose one's will on the community. A moderators' stance is also conditioned by their career stage; experience and seniority in the community.

In many communities, moderation is a collective accomplishment of working through thorny dilemmas. First, there is the question of who should be a moderator and how should this choice be made; this dilemma becomes increasingly salient as a community grows in size. Then, there is the question of labeling behavior as "good" or "bad," and whether norms should be articulated. Lastly, moderators have to reconcile nested and imbricated community norms to accomplish their work, because of the porous community boundaries on Reddit and of global norms enacted by Reddit's administrators.

Community Demand for Controlling Bad Behavior

Contribution Cost and Quality Standards

The context in which an online community exists in will define its norms, users and dictate how its moderators moderate. Communities which cater towards low-cost content will attract low-effort users whom will be moderated by users in a relatively erroneous manner. On the other end of the spectrum, there

[2] All communities are identified by pseudonyms.

are online communities which codify how users should behave, contribute and interact with the community, raising the cost of making a contribution and specifying the content of a valuable contribution. Moderators from stricter online communities tend to favor higher quality content and thus, engage in more discussion around the norms they create than moderators from more relaxed online communities. A community with many norms requires a lot more effort on behalf of the moderators to enact the same moderation actions than a similar community with very few norms:

> *"The rules that we have in place require a focused effort" (AboutWomen Mod.); "We have a higher standard of comment quality. Most subreddits follow, generally, a free-speech type standard, which allows for things like a circlejerk, whereas we want our subreddit to be a carefully curated academic space. [That's a huge] effort [with] over 300,000 people." (AskQuestions Mod.)*

In contrast to the above moderators who have to put a lot of effort into moderation actions, moderators tend to fall back on technical means or on crowd policing to deal with undesirable behavior:

> *"We do more than most users would think, but we're pretty laissez-faire when it comes to moderating. We leave automod to the heavy lifting." (Funny Mod); "Usually, I let the community self-moderate… it creates a better atmosphere than a dictator-like approach." (Townsville Mod)*

Community Visibility and Topic

AboutWomen and DiscussIt, the two communities with the most visibility and polarizing topic, are both more susceptible to bad behavior; their moderators are thus more vigilant. The moderators from other communities made no mention to their susceptibility of being at risk. For instance, the AboutWomen moderator listed the common types of bad behavior which derive from the visible and polarizing nature of the community, which makes the community an easy target for collective forms of bad behavior:

> *"Being a sub that caters to women's opinions opens us up to brigading from the shittier parts of Reddit. In particular, we run into issues with [Male-dominated subreddits] linking to us as smug spectators to our "horrible" opinions. Additionally, some of the "inadequacy" subs will often link when a relevant discussion comes up. What results is a flood of users who display contempt for our rules and opinions." (AboutWomen Mod.)*

Stance toward Controlling Bad Behavior

Orientation among Control Logics

Moderators act as intuitive prosecutors who may draw upon a variety of logics that provide guidance on how to moderate fairly. Some moderators consider that deviants deserve punishment because their focus is on the importance of compliance and adherence to rules while other moderators will consider that deviants can be rehabilitated, by including an understanding of the user's background. For instance, the AboutWomen moderator will consider temporary bans for offenders who she believes can change, and permanent bans for offenders she doesn't believe will change. The DiscussIt moderator will try and appeal to a user's better nature than risk censoring or ban them: *"Appealing to people's better nature sometimes works, and banning is used, rarely."*

Behavior which goes against the stance of a moderator with little tolerance is much more likely to receive a heavy-handed response, regardless of whether it is truly damaging. When undeserved bans or removal of content are handed down, they may garner complaints because the sanctioned user (and the community) perceives the mod's actions as unfair. In some situations, a user will receive an instant ban for behavior that warrants no warning elsewhere. This is not a response to community norms but an outcome of the moderator's inclination toward strong deterrence, especially when it is technically easy to enact.

Another logic which needs to be taken into account concerns the interests of the community and the extent to which content originating from the core, rather than the margins of the community is valued. Newcomers to an online community tend to be sanctioned more severely than established insiders. Newcomers who receive undeserved, negative responses from an online community's members will, therefore, be more likely to leave the community. The AboutWomen moderator recognized this tension: *"Removing the 'shitty' contrary opinions cultivates, to a certain extent, a sort of hive mind where certain opinions are perceived as acceptable and those outside of that are hesitant to contribute."*

Sense of Power

Moderators are aware that they are bestowed with a significant amount of power over their community. Some consider that they must yield this power wisely, by assuming a fiduciary role and self-imposing limitations on their action. However, others reported being happy to yield this power to shape the community as an extension of their self, in an autocratic and arbitrary way at times.

> *"They (the users) see us as gods. We comment and distinguish our names [as moderators] and get showered with ass kissing. [...] Sometimes we feel like kicking up some shit, sometimes we're bored, sometimes it's funny, sometimes a user actually broke a rule. The mods have the ultimate power to make arbitrary decisions. Most users take bans with pride." (Funny Mod.)*

Career Stages in Online Community Maintenance Work

An online community will perceive a moderator as legitimate if they feel that their position is deserved. If a moderator is appointed without any proper selection criteria or is self-appointed, there may be contention between the moderator and the users. Moderators that are newcomers to their role may have to learn the ropes of the community, and establish their credentials before being able to secure the support of the community for their work. The DiscussIt Mod outlined their experience of gaining credibility: *"when I was added as a mod the first time, I was relentlessly downvoted for days because I was not trusted by elements of the community. I have had a varied Reddit experience so that lack of trust was understandable."*

Controlling Bad Behavior as a Collective Accomplishment

Hiring Moderators

Moderators draw (or avoid drawing) a red line between "good" and "bad" behavior, and because this line drawing (or lack thereof) occurs in public, it shapes community norms. As the users in an online community look up to its moderators, they need to represent what an ideal user of the community should be. A moderator who goes against their community's norms can create turmoil. Moderators from communities that grow or suffer from churn have to collectively promote to moderator positions users who fit their stance toward moderation. As these stances are tacit, promotion potential is assessed based on other, more explicit criteria: familiarity with the community, contributions to the community, social ties with moderators, and prior experience moderating. Because stances may vary among moderators of a community, selecting moderators can become an occasion for coalition building, but may also tip the balance of power toward certain views about the governance and purpose of the community.

Labeling Bad Behavior and Articulating Norms

Moderation is a collective accomplishment in the sense that moderators rely upon a jurisprudence record and a community of peers to help in making their choices:

> *"In cases when important players start breaking rules we [are] extra careful reviewing what we have done in the past and following it to the letter. [...] we have a mailing list with all the admins that allow us to discuss things and go over what we want to do [about the issue]" (Gaming Mod.)*

Conflict occurs among moderators, at times because of irreconcilable stances on what should be labeled as "bad" behavior, and what norms should be articulated. All moderators recalled incidents caused by differing stances amongst moderators. In the DiscussIt community, one moderator viewed censorship as unnecessary unless it breached Reddit's site-wide rules and another moderator viewed censorship as a necessary evil. The conflict resulted in one moderator resigning in support of the community goals:

> *"The policy of the sub is not to remove content that breaks the rules of reddit. Bigotry and abuse are implicitly allowed. This is deeply unpleasant, so is used deliberately to sow discord in the subreddit. A troll was pasting anti-Semitic hate speech to the subreddit. One mod removed these comments, I reinstated them. They were removed again, so I simply resigned." (DiscussIt)*

In one community, the conflict escalated to the point where private moderator messages were leaked to the wider community and the moderator involved was removed. Further conflicts arose in the dilemma of whether to remove traces of bad behavior.

Reconciling Nested Community Norms

Because Reddit users flow freely from one community to the next, they may transpose their expectations about what is rightful behavior. What is acceptable to ask or discuss in Funny may be unacceptable in AskQuestions. Newcomers are at particular risk of crossing a red line inadvertently. In these situations, moderators opt for codified norms to dictate behavior. Member fluidity also means that moderators are interlocked in a web of relationships. They may be involved in more than one community at once, as moderators and/or users, making them accountable to audiences who hold conflicting expectations. The AboutWomen Mod described regretting the decision she had made to allow posting of a controversial thread from another subreddit because it had ignited conflict within the team of mods. Those who belonged to two subreddits had felt they were being forced to choose sides. Later there was a lot of back and forth about what they should have been done and should do moving forward.

While a typical online community operates with its own norms and is embedded in societal norms; Reddit communities operate within an additional layer of global norms, codified in its content policy and "reddiquette" (Reddit.com, 2016). Moderators are limited in their ability and capacity to influence the global norms of Reddit, due to a fear of administrators' reprisal in the form of a community shut down. Yet, when asked about clashes between local and global norms moderators had varying motivations for complying with global norms. Some considered the interests of their local community above all else: *"If there was a disagreement with Reddit... held by the entire player base then we would probably move to another platform... in cases of really huge opposition we encourage our players to make their own platforms for discussion if need be." (Gaming Mod.)*

Emerging Contribution & Implications

The key contribution of this emerging research is in providing insights about the situated choices facing mods whose jobs involve maintaining social norms in online communities, and how the enactment of these choices is affected by community demands and the web of relationships in which mods are embedded. By using netnographic techniques, this study complements findings based on experimental, survey, and archival methods about controlling online community behavior (e.g., Kraut & Resnick, 2012). Our findings bring attention to the political and moral struggles that mods face in accomplishing their work, a substantively important yet poorly understood phenomenon. We predict that our findings hold for communities whose purpose is to curate content and where mods are volunteers. Our preliminary findings can be extended by (1) a within-family extension (Barley & Kunda, 2001) to include a broader set of online communities that approximate Reddit's attributes, and (2) by varying moderation parameters to include paid roles and peer production communities. Data collection currently ongoing will provide a broader scope of comparison through additional rich accounts among a greater diversity of Reddit communities.

Some aspects of moderation work are in need of further attention in our theorizing. While controlling others' behavior is at the core of moderation work, the mechanisms that allow how the community and platform owners to control their mods' work are still implicit. While the mods we have interviewed so far feel free to act with impunity (e.g. "*The mods have the ultimate power to make arbitrary decisions.*" – *Funny Mod.*), we suspect that normative and peer control mechanisms (Barker, 1993; Michel, 2011) provide checks on mods' power, but this aspect needs further empirical exploration and theoretical development. Also, given the thorny dilemmas, complex web of relationships, and demands for one's time mods often face, our emergent findings do not yet show how moderation work is accompanied by emotional and physical strain. Online communities, and peer production more broadly, are sometimes considered as a new form of work organization that corrects the flaws of the bureaucratic model of production (Benkler, 2006). Yet, Chen (2009) and Kreiss et al. (2011) warned that work in peer production communities can be liberating as well as constraining. Given the centrality of mods' work in the maintenance of online communities, the resolution of this tension is crucial to the sustainability of vibrant online communities. Lastly, future data collection efforts will need to pay close attention to the role of technology in assisting moderation and curation work. As we have found, technology can be used to alleviate some of the strains felt by mods, by providing affordances to automate tasks, detect deviance, and enforce norms. Yet, studies have also shown that delegating moderation to technology can lead to unintended consequences, such as turning away newcomers and stifling innovation (Halfaker et al., 2012; Kiesler et al., 2012). It thus becomes important to understand the norms regulating technology use by moderators are negotiated in situ, as well as how those norms diffuse through socialization and observation among moderators.

References

Avison, D., & Malaurent, J. 2014. "Is Theory King? Questioning the Theory Fetish in Information Systems," *Journal of Information Technology* (29:4), pp. 327-336.

Barker, J.R. 1993. "Tightening the Iron Cage: Concertive Control in Self-Managing Teams," *Administrative Science Quarterly* (38:3), pp. 408-437.

Barley, S.R., & Kunda, G. 2001. "Bringing Work Back In," *Organization Science* (12:1), pp.76-95.

Barley, S.R. 2015. "Why the Internet Makes Buying a Car Less Loathsome: How Technologies Change Role Relations," *Academy of Management Discoveries* 1(1), pp.5-35.

Bateman, P.J., Gray, P.H. & Butler, B.S. 2011. "Research Note-The Impact of Community Commitment on Participation in Online Communities," *Information Systems Research* (22:4), pp.841-854.

Becker, H.S. 1963. *Outsiders: Studies in the Sociology of Deviance*. New York, NY: The Free Press.

Benkler, Y. 2006. *The Wealth of Networks: How Production Networks Transform Markets and Freedom*. New Haven, CT: Yale University Press.

Bergstrom, K. 2011. "Don't Feed The Troll: Shutting Down Debate About Community Expectations on Reddit.com," *First Monday* (16:8).

Bechky, B.A. 2011. "Making Organizational Theory Work: Institutions, Occupations, and Negotiated Orders," *Organization Science* (22:5), pp.1157-1167.

Binns, A. 2012. "Don't Feed the Trolls! Managing Troublemakers in Magazines' Online Communities." *Journalism Practice* (6:4), pp.547-562.

Bullard, E., & Carter, P. J. 2013. "Effects of Anonymity on Aggression and Hostility on 4chan internet message board: Ethical Issues and Practicalities from a Student Perspective," *The First Annual Cyberpsychology Conference (ACPC)*, De Montfort University, Leicester. Available at http://eprints.hud.ac.uk/18515/

Butler, B., Sproull, L., Kiesler, S., & Kraut, R. 2002. "Community Effort in Online Groups: Who Does the Work and Why," In *Leadership at a Distance: Research in Technologically Supported Work*, Weisband, S. (Ed.), New York, NY: Lawrence Erlbaum Associates, pp. 171-194.

Campbell, J., Fletcher, G., & Greenhill, A. 2009. "Conflict and Identity Shape Shifting in an Online Financial Community," *Information Systems Journal* (19:5), pp.461-478.

Chen, K.K. 2009. *Enabling Creative Chaos: The Organization Behind the Burning Man Event*. Chicago, IL: The University of Chicago Press.

Chesney, T., Coyne, I., Logan, B., & Madden, N. 2009. "Griefing in virtual worlds: causes, casualties and coping strategies," *Information Systems Journal* (19:6), pp. 525-548.

Dibbell, J. 1993, December 21. A Rape in Cyberspace or How an Evil Clown, a Haitian Trickster Spirit, Two Wizards, and a Cast of Dozens Turned a Database into a Society. *Village Voice*, 471-489.

Dokko, G., Nigam, A., & Rosenkopf, L. 2012. "Keeping steady as she goes: A negotiated order perspective on technological evolution," *Organization Studies* (33: 5-6), pp. 681-703.

Gioia, D., Corley, K., & Hamilton, A. 2013. "Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology," *Organizational Research Methods* (16:1), pp. 15-31.

Goldspink, C. 2010. "Normative Behaviour in Wikipedia," *Information, Communication & Society* (15:5), pp.652-673.

Halfaker, A., Geiger, R.S., Morgan, J.T., & Riedl, J. 2012. "The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline," *American Behavioral Scientist* (57:5), pp. 664-688.

Hallett, T., Shulman, D., & Fine, G. A. 2009. "Peopling Organizations: The Promise of Classic Symbolic Interactionism for an Inhabited Institutionalism," In *The Oxford Handbook of Sociology and Organization Studies,* Adler, P. (Ed.), Oxford, UK: Oxford University Press, pp. 486-509.

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for Safety Online: Managing 'Trolling' in a Feminist Forum," *The Information Society* (18:5), pp.371-384.

Henfridsson, O. 2014. "The Power of an Intellectual Account: Developing Stories of the Digital Age," *Journal of Information Technology* (29:4), pp. 356-357.

Kazmer, M.M., & Xie, B. 2008. "Qualitative Interviewing in Internet Studies: Playing with the Media, Playing with the Method," *Information, Communication & Society* (11:2), pp.257-278.

Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. 2012. "Regulating Behavior in Online Communities," In *Building Successful Online Communities: Evidence-Based Social Design*, Kraut, R., Resnick, P. (Eds.), Cambridge, MA: MIT Press, pp. 125-178.

Kling, R. 1978. "Patterns Of Segmentation And Intersection In The Computing World," *Symbolic Interaction* (1:2), pp.24-43.

Kollock, P., Smith, M. 1994. "Managing the Virtual Commons: Cooperation and Conflict in Computer Communities," In *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*, Herring, S.C. (Ed.) Amsterdam, Netherlands: John Benjamins.

Kozinets, R.V. 2015. *Netnography: Redefined*. Thousand Oaks, CA: Sage Publications.

Kraut, R.E., Resnick, P. 2012. *Building Successful Online Communities: Evidence-based Social Design*. Cambridge, MA: MIT Press.

Kreiss, D., Finn, M., & Turner, F. 2011. "The Limits of Peer Production: Some Reminders from Max Weber for the Network Society," *New Media & Society* (13:2), pp. 243-259.

Lamb, R., & Kling, R. 2003. "Reconceptualizing users as social actors in information systems research," *MIS Quarterly* (27:2), pp.197-236.

Lampe, C. and Johnston, E., 2005, November. "Follow the (Slash) Dot: Effects of Feedback on New Members in an Online Community," In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, pp. 11-20.

Lampe, C., & Resnick, P. 2004. "Slash (Dot) and Burn: Distributed Moderation in a Large Online Conversation Space," In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 543-550.

Lampe, C., Zube, P., Lee, J., Park, C., & Johnston, E. 2014. "Crowdsourcing Civility: A Natural Experiment Examining the Effects of Distributed Moderation in Online Forums," *Government Information Quarterly* (31:2), pp.317-326.

Leonardi, P. M. 2012. Materiality, Sociomateriality, and Socio-technical Systems: What Do These Terms Mean? How are They Related? Do We Need Them? In *Materiality and Organizing: Social Interaction in a Technological World*, Leonardi, P.M., Nardi, B.A., Kallinikos, J. (Eds.), Oxford, UK: Oxford University Press, pp. 25-48.

Magnusson, J., & Nilsson, A. 2013. "Introducing app stores into a packaged software ecosystem: a negotiated order perspective," *International Journal of Business Information Systems* (14:2), pp. 223-237.

Markus, M.L. 2014. "Maybe not the King, but an Invaluable Subordinate: A Commentary on Avison and Malaurent's Advocacy of 'Theory Light' IS Research," *Journal of Information Technology* (29:4), pp. 341-345.

Michel, A. 2011. "Transcending Socialization: A Nine-Year Ethnography of the Body's Role in Organizational Control and Knowledge Workers' Transformation," *Administrative Science Quarterly* (56:3), pp. 325-368.

Poor, N. 2005. "Mechanisms of an Online Public Sphere: The Website Slashdot," *Journal of Computer-Mediated Communication* (10:2), pp. 00-00.

Reddit.com. (2016). Reddit Content Policy. Retrieved May 2, 2016, from https://www.reddit.com/help/contentpolicy

Strauss, A., Schatzman, L., Ehrlich, D., Bucher, R., & Sabshin, M. 1963. "The Hospital and its Negotiated Order," In *The Hospital in Modern Society*. Friedson, E. (ed.) Glencoe, NY: The Free Press, pp. 147–169.

Vadera, A., Pratt, M., & Mishra, P. 2013. "Constructive Deviance in Organizations: Integrating and Moving Forward," *Journal of Management* (39:5), pp. 1221-1276.

Warren, D. 2003. "Constructive and Destructive Deviance in Organizations," *Academy of Management Review* (28:4), 622-632.